Enabling Bayesian Inference for the Astronomy Masses

Performance Report submitted by M. D. Weinberg, PI
Period: 3/15/07–3/14/08
Grant Number: NNG-06-GF25G

1 Executive summary

Active development proceeded in four of the five defined research topics:

1. Development of statistical methodology
    - Implemented Gelman-Rubin parallel chain tests along with a vanilla *Parallel Chains* class for comparison with tempered schemes
    - Implemented generic tempered simulation with a generalized concept of "hot" chains beyond powering-up the posterior distribution.
    - Implemented and tested a posterior distribution representation based on Metric Tree kernel density estimation.
    - Preliminary implementation of an equi-energy sampler.
2. Development of persistence technology
    - Re-implemented of serialization and persistence methods based on the standard Boost C++ libraries. design based on preprocessor macros
    - Preliminary tests of serial and parallel versions of the persistence package has been tested.
    - Developed detailed plan for persistent store based on SVN repositories
3. Astronomical applications
    - Developed and implemented a GALFIT-like galaxy image analyzer which we have code-named *GALPHAT* for GALaxy PHotometric ATtributes. We found numerous problems in the commonly used package GALFIT that necessitated a full re-implementation.

- Developed and implemented semi-analytic method routine. We are in process of characterizing the probabilistic properties of the semi-analytic-method (SAM) process. Mixing and convergence are challenging with SAMs, as anticipated.
- We have implemented a color-magnitude diagram *generator* based on the most recent isochrone tracks that included the AGB and post-AGB stellar evolution phases. We may now make start count predictions along any line of sight given a star-formation history, metallicity distribution in space. We will test both our model and methodology using 2MASS LMC/SMC data before beginning to apply this to galactic problems. This work was led by post-doc Jörg Colberg.

I will detail some of advances below and end with a list of Milestones for Year 3.

2 Research milestones and summary

1 Persistence subsystem development

This year we completed the conversion of the persistence subsystem of the software (save and restore of ongoing work). We replaced our home-brewed prototype implementation with one using the well-documented and supported Boost library for persistence. As we did this, we simplified what the programmer writes in terms of annotations inserted into the code for C++ classes, and replaced the scripts that process annotated code with new, more reliable scripts. While this work could all be considered just good maintenance, it laid the ground for us to add checkpointing.

Checkpointing goes beyond save/restore in that it saves what is happening in the *middle* of a Markov chain. Running these chains is the most time-consuming part of running the BIE, and thus the most vulnerable to crashes, etc. Checkpointing allows one to resume after a crash, or if one needed or desired to abandon a computation for some reason. We can trigger checkpointing based on the number of iterations since the last checkpoint, the amount of time that has passed, or upon user request via typing a certain control character on the console.

Finally, we have developed a design for our “lab book” extension of the software. This will record commands and allow one to go back to any point in the sequence and branch off in a different direction, with all the exploration paths *saved persistently* and labeled for possible future use. Among other things, this supports both good organization of experiments and results and provides strong evidence of the provenance of those results. We reviewed various underlying libraries that could support this work and settled on subversion, a widely available source code versioning system, that also support versioning of binary data and (important for our application) a standard library interface. It is also stable, reliable, and well-maintained, and likely to remain so.

2 GALPHAT

2.1 Motivation

The galaxy structure is evolving due to gravitational and gas dynamical physics in the expanding Universe. To understand the evolution of galaxy structure based on their morphology has been done by human eye, which led to the systems in use today such as Hubble type. As galaxy surveys have become deeper and more voluminous, researchers have explored a variety of automatic classification schemes.

We had originally intended to use BIE as a backend for GALFIT . GALFIT is a modular package written to perform two dimensional image decompositions for galaxies which are from nearby to distant (Peng et al., 2002). GALFIT takes an input image and outputs a model-subtracted images as well as a catalog of structural parameters for an arbitrary number of components. Each predefined component has up to ten parameters but allows for an arbitrary number of user-defined profiles and components. Some parameters may be fixed depending on one's application but a typical fit will require greater than 12 parameters. We found that the pixel integration and PSF convolution were too inaccurate and time-consuming for our application which necessitated our rewriting the model generation code. We have code-named the new parameter determination package GALPHAT for GALaxy PHotomometric ATtributes. Our combination of this approach with our Bayesian Inference Engine back end, which will allow GALFIT-based investigations of the full posterior not just the extremum mode, and will establish proper prior distributions, which allow inferences using Bayes Factors over a wide variety of competing models and hypotheses.

2.2 Bayesian approach for modelling data

For the likelihood function, we construct the likelihood function using models in GALFIT .

$$P(D \mid \theta) = \frac{exp(-\frac{1}{2}[\mathbf{D}-\mathbf{M}(\theta)]^t\mathbf{W}[\mathbf{D}-\mathbf{M}(\theta)])}{(2\pi)^{Npix/2} \mid \mathbf{W} \mid^{-1/2}} \tag{1}$$

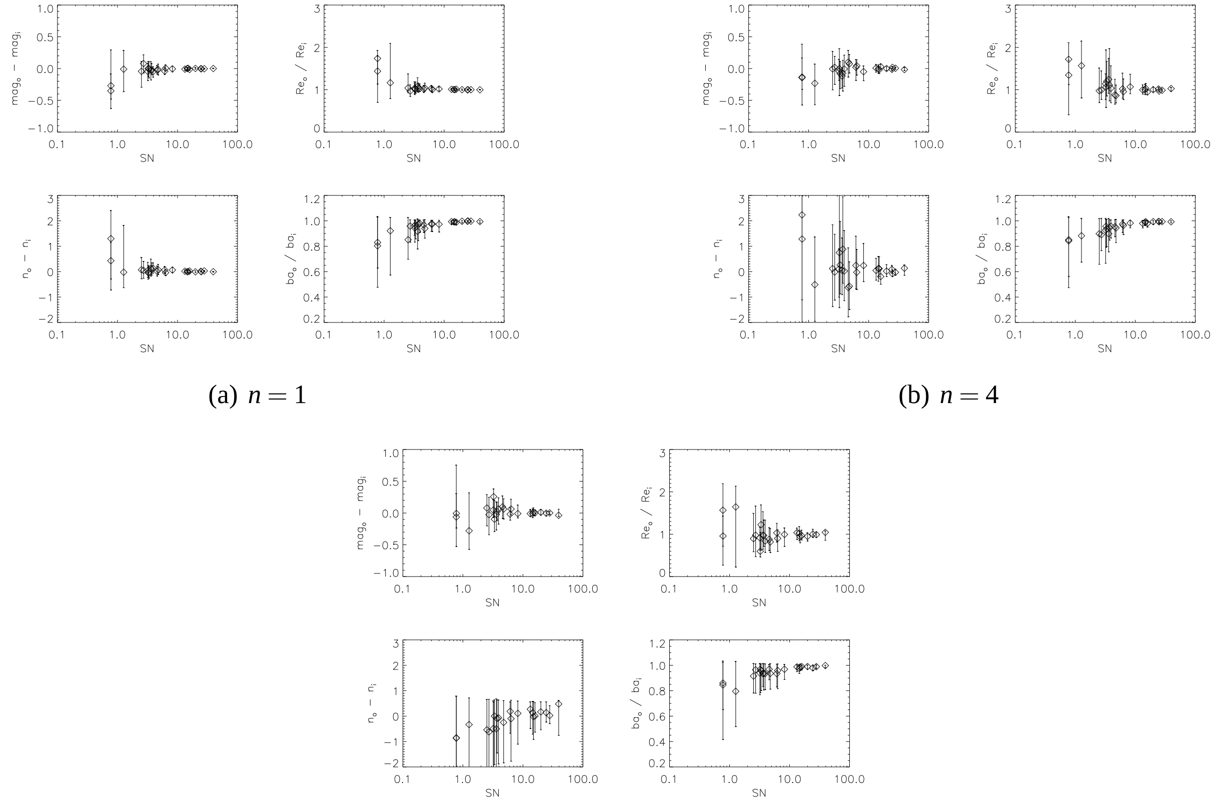
where **D** is data vector($N_x \times N_y$), $\mathbf{M}(\theta)$ is a model vector and **W** is a weight matrix for pixel value.

For the prior for parameters, we mostly adopt the uniform prior with a range(top-hat) which leads the likelihood dominated posterior probability distribution and basically the same case with the maximum likelihood method, a least informative case of Bayesian statistics. As we shall see in later, the effect of prior becomes more significant when we have data where the information is weak and degenerated. For example, in case of low *S/N* data, the informative priors for some parameters help to obtain the robust estimate for those parameters.

We will give examples below of GALPHAT application in synthetic tests and to 2MASS images.

2.3 Ensemble examples

These examples investigate bias and confidence as a function of signal to noise ratio (S/N). Here, S/R is defined as the ratio of photons inside the half-light radius to that of the background inside of the same radius. The Figure 1 describes the parameter distributions for Sésic indices $n = 1, 4, 7.5$. Each galaxy has same half-light radius of 10 and axis ratio of 1. The figures show the residual (for magnitude, Sérsic index) or ratio(for half-light radius and axis ratio) of model parameter with respect to the true input parameter as a function of mean SN. We use 20,000 converged samples and perform kernel density estimation for estimating the probability density of parameter. The median value of the estimated parameter is shown (diamond symbol) and the error bars describe the 99.7% confidence interval.

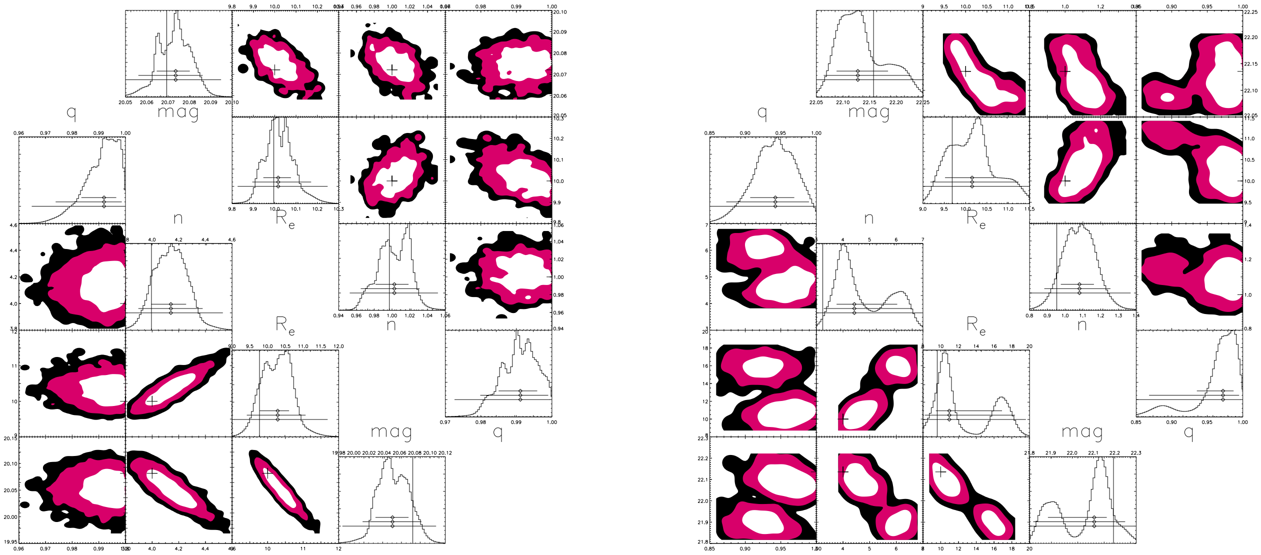


(a) $n = 1$

(b) $n = 4$

(c) $n = 7.5$

Figure 1: Sérsic models parameter estimation for a variety of values S/N along the ordinate. Each sub panel shows the recovered value of the magnitude, effective radius R_e, axis ration b/a and index. The diamond symbol indicates and the error bars the 99.7% confidence interval.



(a) S/N=15 (b) S/N=3

Figure 2: One dimensional marginalized parameter posterior and two dimensional marginalized joint distribution of model parameters for 4 galaxies from the above sample shown in Figure 1 for two values of S/N: 15 and 3. The marginalized distribution of magnitude, half-light radius, Sérsic index and axis ratio are shown and contours of their joint distribution are represented 68.5, 95.4 and 99.7% confidence intervals. In 1D marginalized plot, symbols are sample median and bars are corresponding same confidence interval as the contour level.

One dimensional marginalized parameter posterior and two dimensional marginalized joint distribution of model parameters for 4 galaxies from the above sample shown in Figure 1 are shown in Figure 2. In this matrix plot, ”upper off-diagonal” part is for synthetic galaxy with Sérsic index, 1.0 and ”lower off-diagonal” part is for synthetic galaxy with Sérsic index, 4.0. Image size is 200 by 200 and typical wall clock time with AMD Athlon MP 1800+, 1.5GHz is 2 hours for getting 20,000 samples.

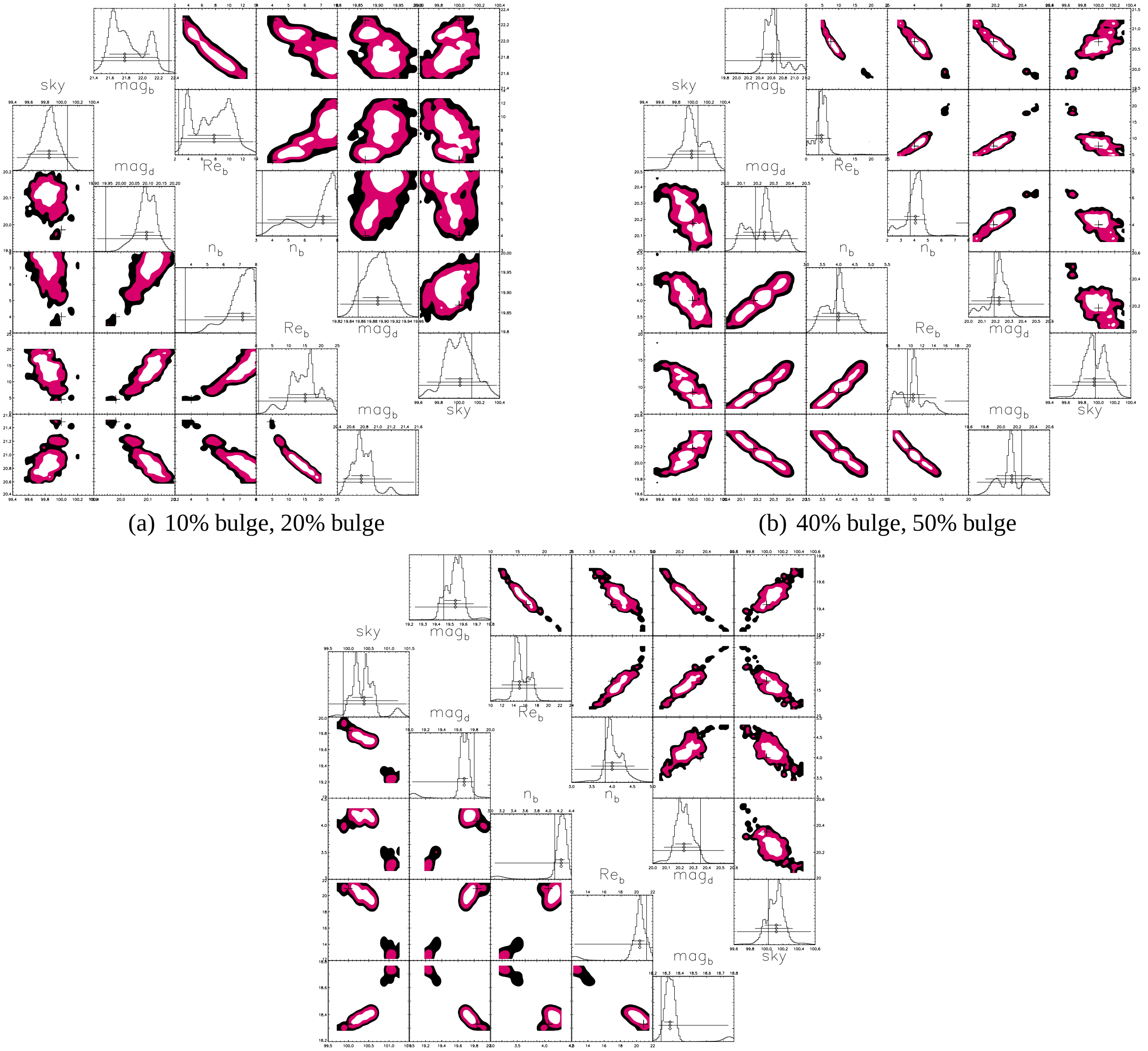
2.4 2-component-model examples

Here we describe the results of similar experiments for two component(bulge + disk) synthetic galaxies. We generated many galaxies with different bulge to total light ratio (B/T) from 0.1 to 0.8. Figure 3. Again, we used 40,000 samples with an image size is 200 by 200 and typical wall clock time with AMD Athlon MP 1800+, 1.5GHz is 7 hours.

Figure 4 shows parameter distribution (c.f. Fig. 1 for the two-component models. This is the same kind of figure as in 1, but for two components(bulge+disk). I showed residual / ratio of parameters against to the true input parameters.

2.5 NGC 137, NGC 311, NGC 470

These are examples of 2MASS galaxy image analyses in K band($\sim$ 10 mag). 40,000 samples are used for the parameter estimations described in Figure 5.



(a) 10% bulge, 20% bulge

(b) 40% bulge, 50% bulge

(c) 70% bulge, 80% bulge

Figure 3: Marginalized distribution of bulge magnitude, bulge half-light radius, bulge Sérsic index, bulge axis ratio, disk magnitude, disk half-light radius, disk axis ratio and sky background (see Fig. 2. Each panel describes two bulge fractions. E.g. in the first panel the upper off-diagonal part for a 10% bulge fraction and lower off-diagonal part for a 20% bulge fraction.

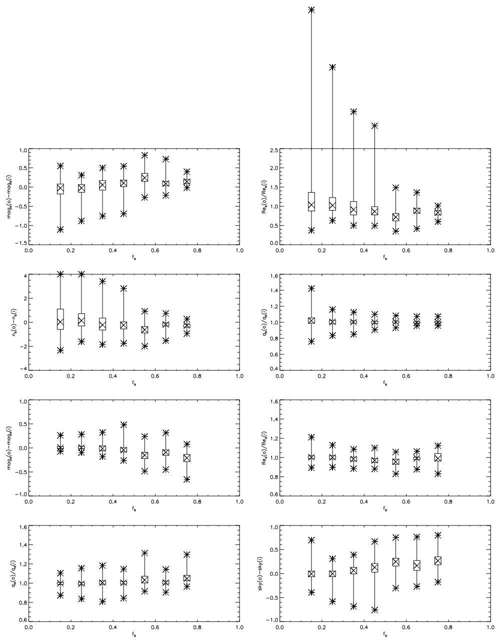


Figure 4: Inferred distribution of bulge magnitude, half-light radius, Sérsic index, axis ratio disk magnitude, half-light radius, axis ratio and sky background for the two-component models as function of bulge fraction. Each bulge fraction bin is generated 10 galaxies with slightly different size and range of axis ratio/PA. All disk components have mean SN=3 and bulge SN spans from ~ 12 to ~ 27. The box and whisker symbols are as follows: the box covers from 25% to 75% quantile of ensemble of data from 10 galaxies and whiskers are min and max of data. Symbols in each box correspond median of data.

2.6 Conclusions

1. GALPHAT recovers input parameters for modest signal-to-noise ratio with little bias and performs remarkably well for low signal-to-noise ratio.

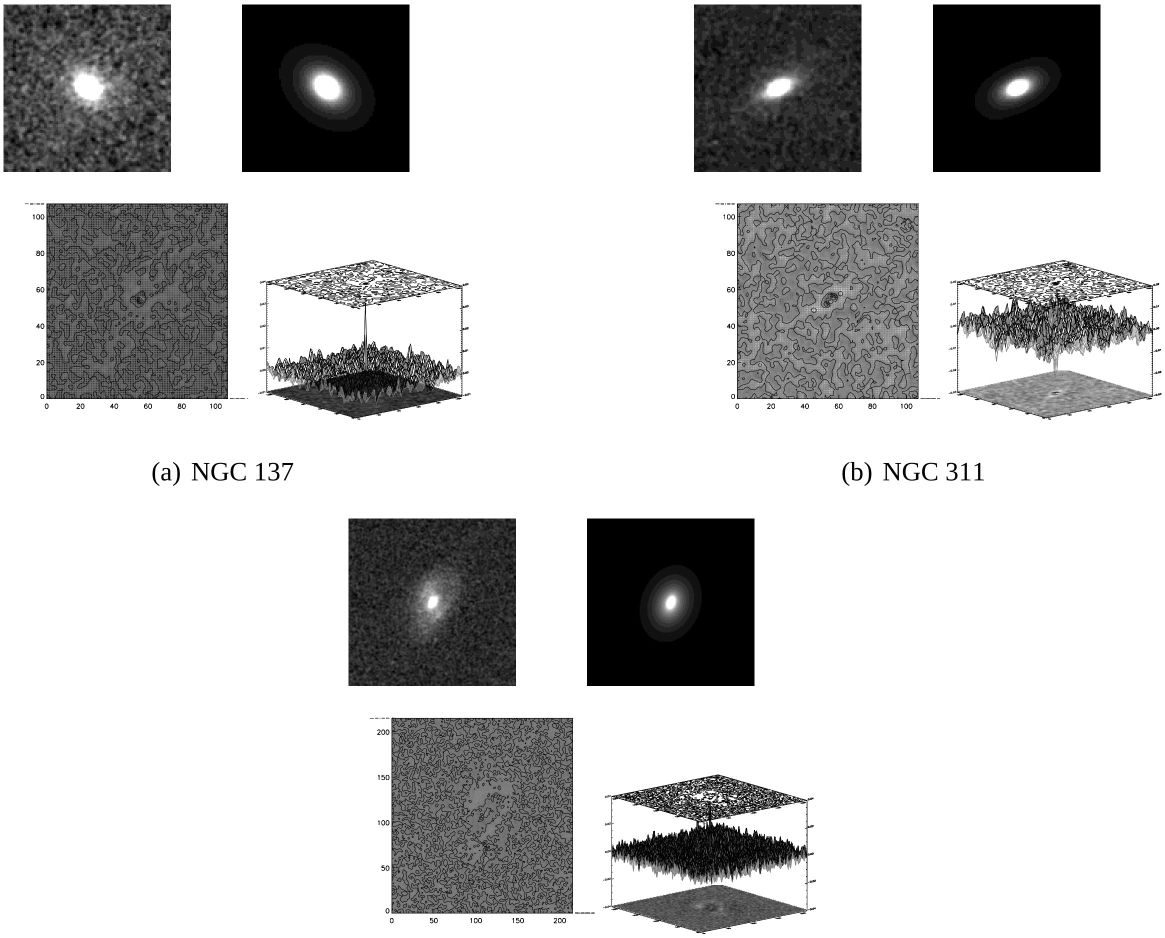
2. Similarly, multiple component inferences correctly do not reject input parameters in the presence of degeneracies, an important and satisfying confirmation of the methodology.

3. Models may be productively compared using *Bayes Ratios*.

4. Preliminary tests show that GALPHAT parameter determination agrees with GALFIT analyses, converge quickly, and provide robust confidence intervals from the full posterior distribution while diagnosing degeneracies.

3 SAMS–BIE

Semi-Analytic Models (SAMs) have been extensively used to study the formation and evolution of galaxies (e.g. Kauffmann et al., 1999; Somerville and Primack, 1999; Cole et al., 2000). In SAMs, one starts with a catalog of merger trees which describe the assembly of individual dark matter halos, and all other additional physical processes, e.g. gas cooling, star formation and feedback, AGN, galaxy mergers, etc., are also added into SAMs through empirical functions. As previously described, we have developed sophisticated programs to generate the merger trees using Monte-Carlo methods. For a given halo mass M_2 at a given redshift z_2, we calculate the conditional probability for such a halo having a progenitor with with mass $M_1 < M_2$ at an earlier redshift z_1.



(a) NGC 137

(b) NGC 311

(c) NGC 470

Figure 5: Parameter estimation for 2MASS K-band images as labeled. Each panel shows the image (upper left) and model (upper right), along with the absolute residual image (lower left) and an edge-on wire-frame representation of relative residual image (lower right). In general, residual is less than 1% but spiky feature in the central region show maximum $\sim$ 4% relative difference in the case of NGC1 470.

We generate random numbers according to the conditional probability to allocate progenitor halo masses.

In the binary tree without accretion, at each time step a halo either splits into two progenitors or does not fragment but retain its mass. In practice, to make the Monte-Carlo method more efficient, we change variables. Instead of redshift and mass, we choose $\omega \equiv \delta_c(z) = \delta_{c,0}/D(z)$ as our time variable, and $S(M) \equiv \sigma^2(M)$ as our mass variable. The probability for taking a new step ΔS in a time-step $\Delta\omega$ is

$$P(\Delta S, \Delta\omega) \mathrm{d}\Delta S = \frac{1}{\sqrt{2\pi}} \frac{\Delta\omega}{(\Delta S)^{3/2}} \exp\left[-\frac{(\Delta\omega)^2}{2\Delta S}\right] \mathrm{d}\Delta S. \tag{2}$$

If we make a change in variables further, $x \equiv \Delta\omega/(2\sqrt{\Delta S})$, the variable x becomes a Gaussian distribution with zero mean and unit variance. By generating a Gaussian random number, we produce a new mass for one of the two progenitor halos and the rest mass if any is assigned for the other progenitor.

3.1 Current results

We have found that our Bayesian SAM model, BIE-SAM convergences quickly for a small number of free parameters but poorly for a large number of free parameters typical of real-world SAMs. This behavior was anticipated, motivated the proposed application, and further underlines the importance of sound probabilistic methodology for cosmological and galaxy formation applications. These difficulties have motivated us to collaborate with a Michael Lavine (probabilist/statistician formally at Duke now at UMass) to develop hybrid MCMC schemes appropriate for our high-dimensional, poor-mixing situation.

We illustrate degenerate and under-constrained SAMs from converged BIE-SAM run with 8 free parameters[1] The free parameters are

1. the cut-off halo mass for radiative cooling,
2. the amplitude for the star formation efficiency law,
3. the power index for the star formation efficiency halo mass dependence law,
4. the total supernova feedback energy fraction
5. the amplitude for supernova reheating law
6. the power index for the supernova reheating halo mass dependence law,
7. the fraction of the reheated gas ejected
8. the galaxy merging timescale in terms of the dynamical friction timescale (equivalent to $1/\ln\Lambda$).

[1] “Real-world” SAM problems often have > 20 parameters.

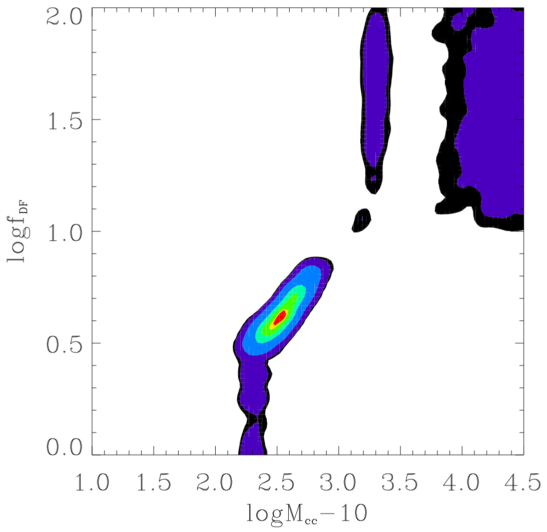


Figure 6: The marginalized posterior on the cooling cut-off halo mass and the merging timescale plain. The color coded contours correspond to 95, 90, 70, 50, 30, 10, and 5 percent confidence level.

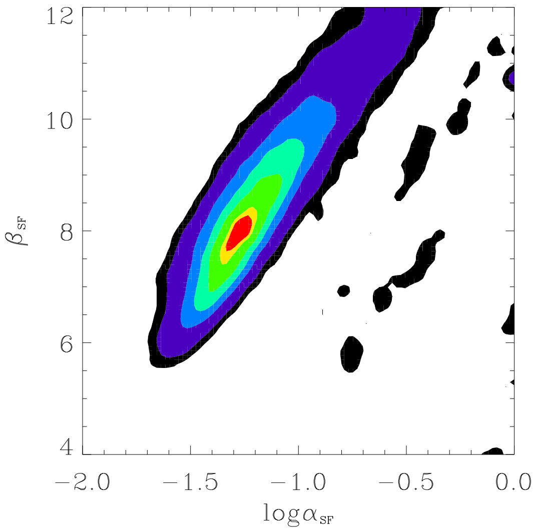


Figure 7: The marginalized posterior on the amplitude of the star formation efficiency and the power index of the star formation efficiency halo mass dependence law. These two parameters are mildly degenerate under the constrain of the stellar mass function.

We use the observed stellar mass function to constrain the model. Some parameters, cooling cut-off and merging timescale for instance, are well constrained, but some parameters show degeneracy, and some parameters are not constrained at all. We used differential evolution MCMC algorithm with 128 chains for more than 6000 iterations. The posterior with 2 outliers removed satisfies the Gelman-Rubin test with ($\hat{R} < 1.2$). Figures 6–10 describe the results.

4 Star count analyses

With deep data sets of asymptotic giant branch (AGB) stars from both the Large and the Small Magellanic Cloud (LMC and SMC, respectively), it should be possible to model the structure of these galaxies using theoretical models for such stars. The LMC and SMC provide a nice laboratory for such studies, especially since galactic extinction towards the Clouds is extremely small and AGB stars are very bright, so the galaxies are well resolved observationally. As input data (prior) for the Bayesian Inference machinery, color–mmagnitude diagrams (CMDs) have to be produced, which are based on observed and theoretical stellar properties and on a model of the structure of the galaxies.

4.1 Generating CMDs from Isochrones

Up until now, the work has focused on generating CMDs from sets of theoretical isochrones, available at five different metallicities, which are described in Cioni et al. (2006). Generating these

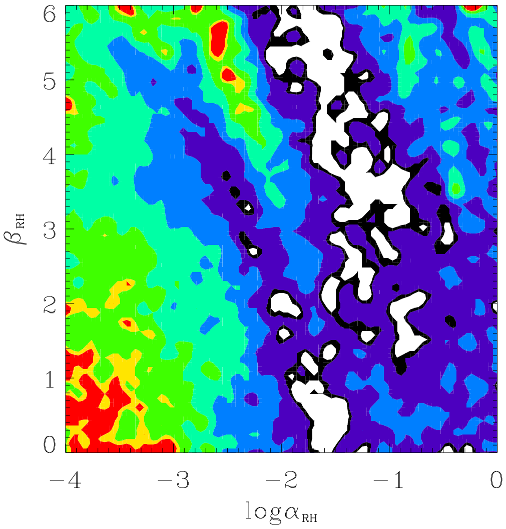


Figure 8: The marginalized posterior on the amplitude of supernova reheating and the power index of the supernova reheating halo mass dependence law. These parameters are not well constrained by the stellar mass function.

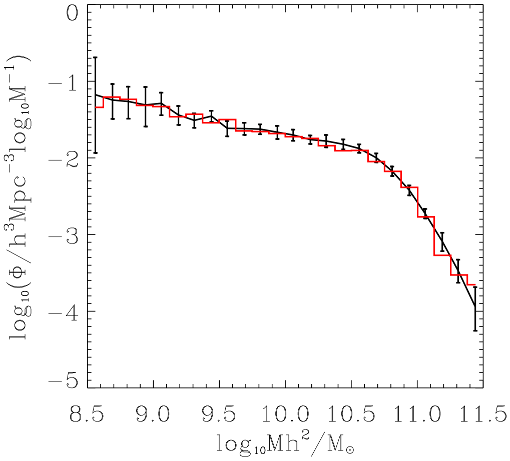


Figure 9: The predicted stellar mass function (red) by a parameter set randomly selected within the 10 percent confidence region compared with observation (black). The flat faint-end and the steep massive-end are well reproduced.

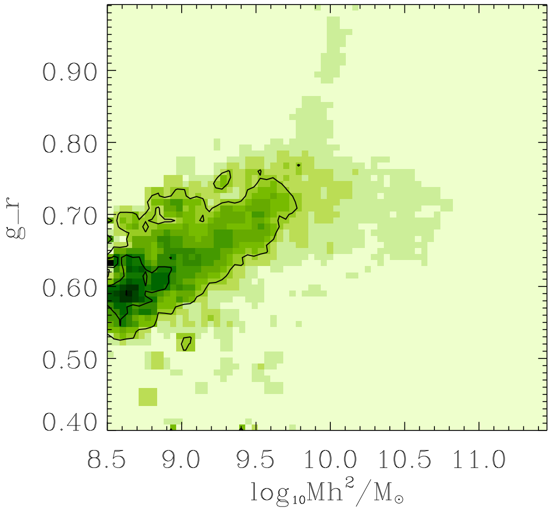


Figure 10: The predicted $g-r$ color-magnitude diagram with the same model parameter set. The diagram is far from observations.

CMDs requires a sequence of interpolations (in stellar mass, metallicity, and age) to produce a realistic input–data set (the prior) for the Bayesian Inference machinery.

Apart from the isochrones, star–formation rate (SFR) histories (SFRHs) and age–metallicity relations (AMRs) have to be provided – these are take from Pagel & Tautvaisiene (1998) and especially the more detailed and recent work of Carrera et al. (2008).

Given the shapes of the isochrones of the stars of interest here, they had to treat them with special care, in order to ensure that the interpolations between neighboring isochrones would use the same evolutionary stages of stars. This process proved to be extremely tedious and time–consuming as it could only be automated in part and required inspection by hand for each of the isochrones.

In addition to an SFRH and AMR, an initial mass function (IMF) has to be provided. Following Cioni et al. (2006), we assume it to be independent of age and equal to the log–normal function of Chabrier (2001)[2].

4.2 Example CMDs

In Figure 11, we show four simulated CMDs for four different models. The top row contains simple toy models, which assume a fixed metallicity ($Z = 0.008$) and exponentially increasing (panel a) and exponentially decreasing SFRHs (panel b).

The bottom row shows two realistic models, both of which use input data based on observational data of the SFRH and on theoretical modeling of the AMR. Panel c uses the observed SFRH (Carrera et al. 2008, their Figure 17) and the bursting–model AMR by Pagel & Tautvaisiene (1998) for an LMC bar field. Panel d uses the observed SFRH (Carrera et al. 2008, their Figure 17) and the closed–box model ($y = 0.008$) by Carrera et al. (2008; see their Figure 18) for an LMC disk field.

3 Milestones for Year 3

1. Statistical & MCMC development
   Continued testing and exploration of novel techniques for rapid improvement of mixing and convergence for high-dimensional complex posterior distributions typical of real-world astronomical problems.

   We hope to provide qualitative suggestions and wisdom for choosing various MCMC algorithms and diagnostic procedures. A paper describing the features and use of the BIE is in preparation.

2. Persistence subsystem

   We anticipate a working implementation of our persistence subsystem by the end of July 2008. This will support recording computations and the relationships between inputs and outputs, in a research log, so that one can always go back and determine the origin of data and how it was processed, replaying from a previous state, but with different commands or

[2] And just like Cioni et al. (2006) we find that the LMC CMDs generated with our code do not depend on the detailed shape of the IMF.

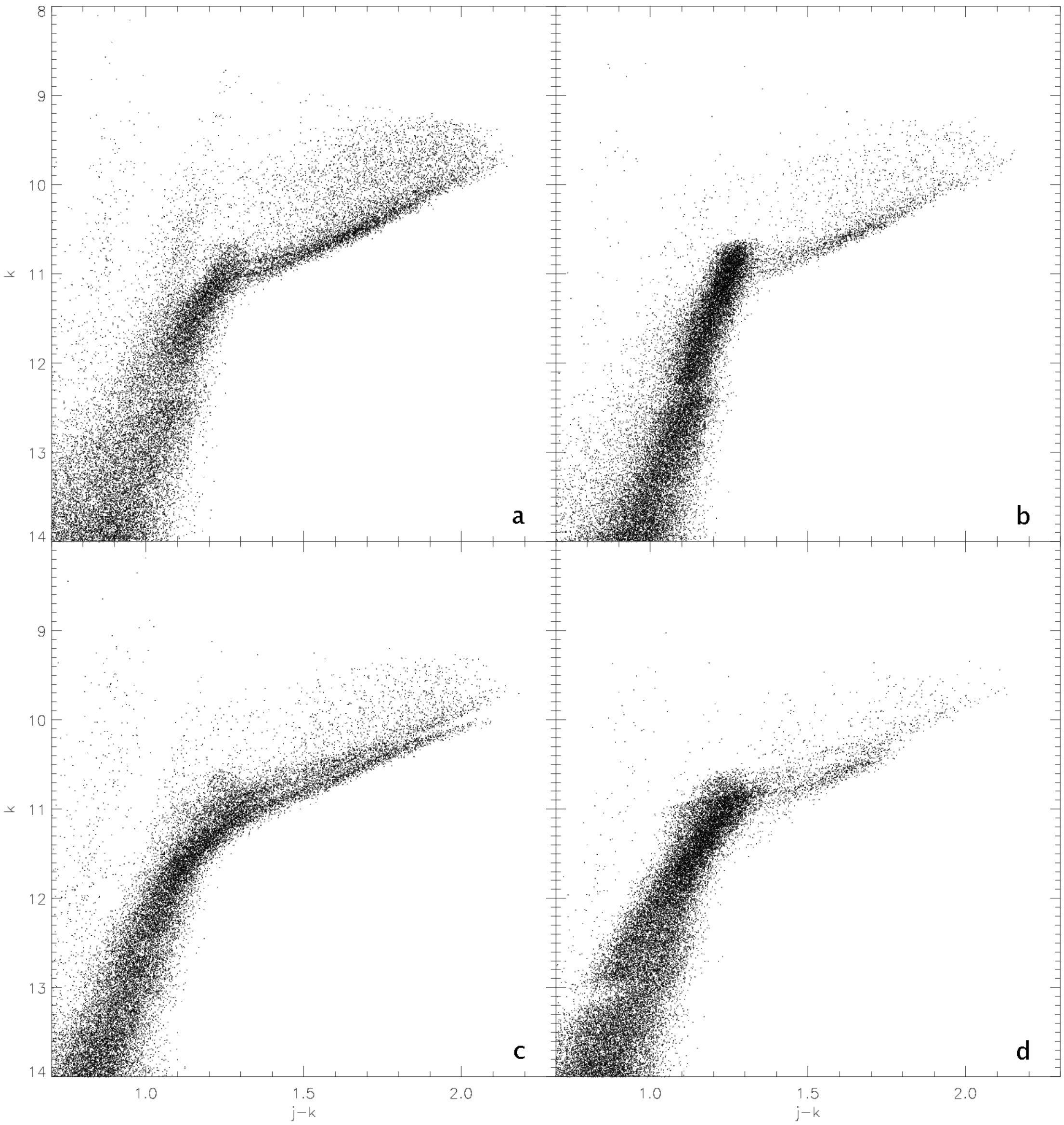


Figure 11: Simulated CMDs for different models. Top row: Fixed metallicity $Z = 0.008$ (constant AMR) with exponentially increasing SFR (panel a), exponentially decreasing SFR (panel b). Bottom row: Assuming observed SFRH (Carrera et al. 2008) and bursting–model AMR of Pagel & Tautvaisiene (1998) for LMC bar field (panel c) and observed SFRH (Carrera et al. 2008) and $y = 0.008$ closed–box model AMR from Carrera et al. (2008) for LMC disk field (panel d).

parameters—what we call *what-if* exploration. One can always go back to some previous time or step and compute forward in new directions, and checkpointing and recovery.

3. BIE–GALFIT

   We are currently testing idealized data sets and benchmarking the efficiency of BIE in hypothesis testing. During Year 3, we anticipate moving on to inference on real astronomical data and publications demonstrating the methods and application. In addition, we are currently implementing computational optimizations that will allow production analysis. We anticipate a full-up stand-alone version of BIE to be released to the public in the upcoming year. Two papers are currently in preparation; we anticipate two more in the upcoming year.

4. Semi-analytic models

   We will continue to improve the performance of our SAM implementation and test its performance and develop hybrid MCMC algorithms to expedite mixing. In Year 3, we plan to apply the Bayes Factor methodology to test specific hypotheses about the importance of various parameters in the underlying physical mechanisms or used to test the effect of different physical hypotheses, i.e., different parameterizations and combinations of physical processes, without the constraint that their prescriptions be nested. We have begun discussions with other SAM practitioners hope to test BIE with their codes as well. We have two papers in the planning outline stage and anticipate submission by the end of the calendar year 2008.

5. Star-count analyses

   Writing and testing the code to generate model CMDs is almost complete, with only minor further testing required (this part of the project will be finished before the end of July 2008).

   As a next step, the code to generate the CMDs will have to be incorporated into the existing Bayesian Inference Estimation (BIE) code. There already exist modules in that code that produce very simple CMDs, and adding the new code to the BIE code should be straightforward, hopefully requiring only minor work. We expect this work, including the necessary test phase, to be done by the end of August 2008.

   At that stage, the machinery is production ready, and we will start to model the structure of the LMC, anticipating first results by mid–October 2008. Given applying the code to other galaxies requires no additional work other than modeling those galaxies' structure, we will extent the LMC work to the SMC and, very possibly, to other close–by galaxies, to conclude these studies by early 2009.

Bibliography

Carrera, R., Gallart, C., Hardy, E., Aparicio, A., and Zinn, R. 2008, AJ, 135, 836.
Chabrier, G. 2001, ApJ, 554, 1274.
Cioni, M.-R. L., Girardi, L., Marigo, P., and Habing, H. J. 2006, A&A, 448, 77.
Cole, S., Lacey, C. G., Baugh, C. M., and Frenk, C. S. 2000, MNRAS, 319, 168.
Kauffmann, G., Colberg, J. M., Diaferio, A., and White, S. D. M. 1999, MNRAS, 303, 188.
Pagel, B. E. J. and Tautvaisiene, G. 1998, MNRAS, 299, 535.
Peng, C. Y., Ho, L. C., Impey, C. D., and Rix, H.-W. 2002, AJ, 124, 266.

Somerville, R. S. and Primack, J. R. 1999, MNRAS, 310, 1087.